The past, present, and future of `data.table`

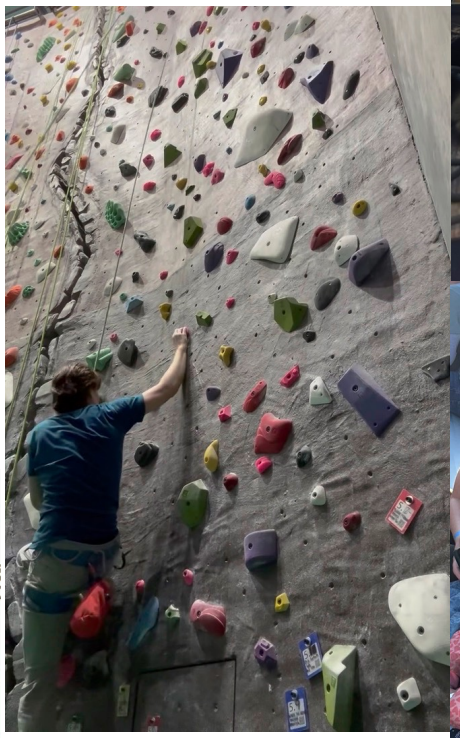
Tyson S. Barrett, PhD | useR! 2024 | Salzburg, Austria

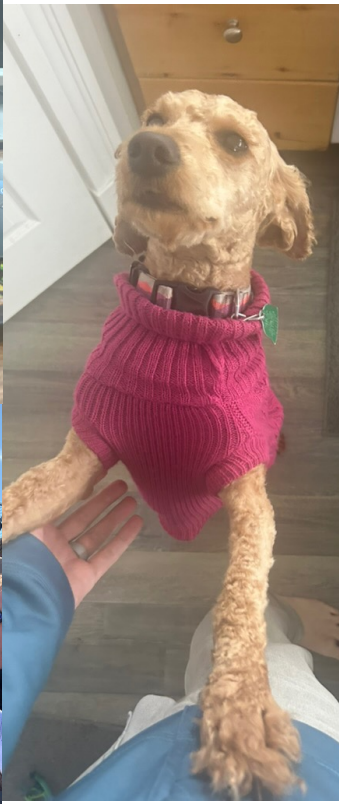
WHO AM I?

Current maintainer of `data.table`

Research Manager at a large US Health Insurer

Human nerd





WHY data.table?

```
dt[i, j, by]
```

Concise syntax

Fast speed

Memory efficient

Careful API lifecycle management

Community

Feature rich

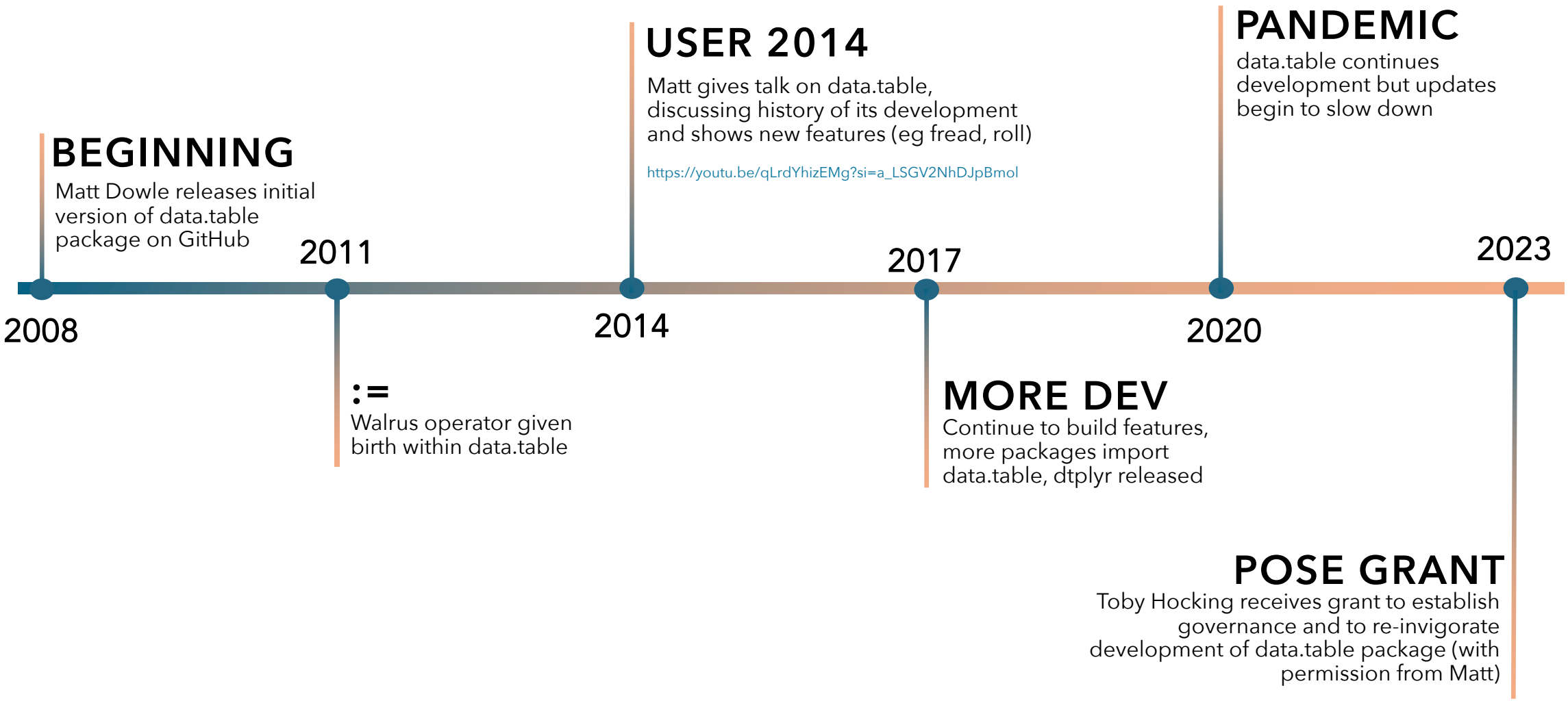
WHY data.table?

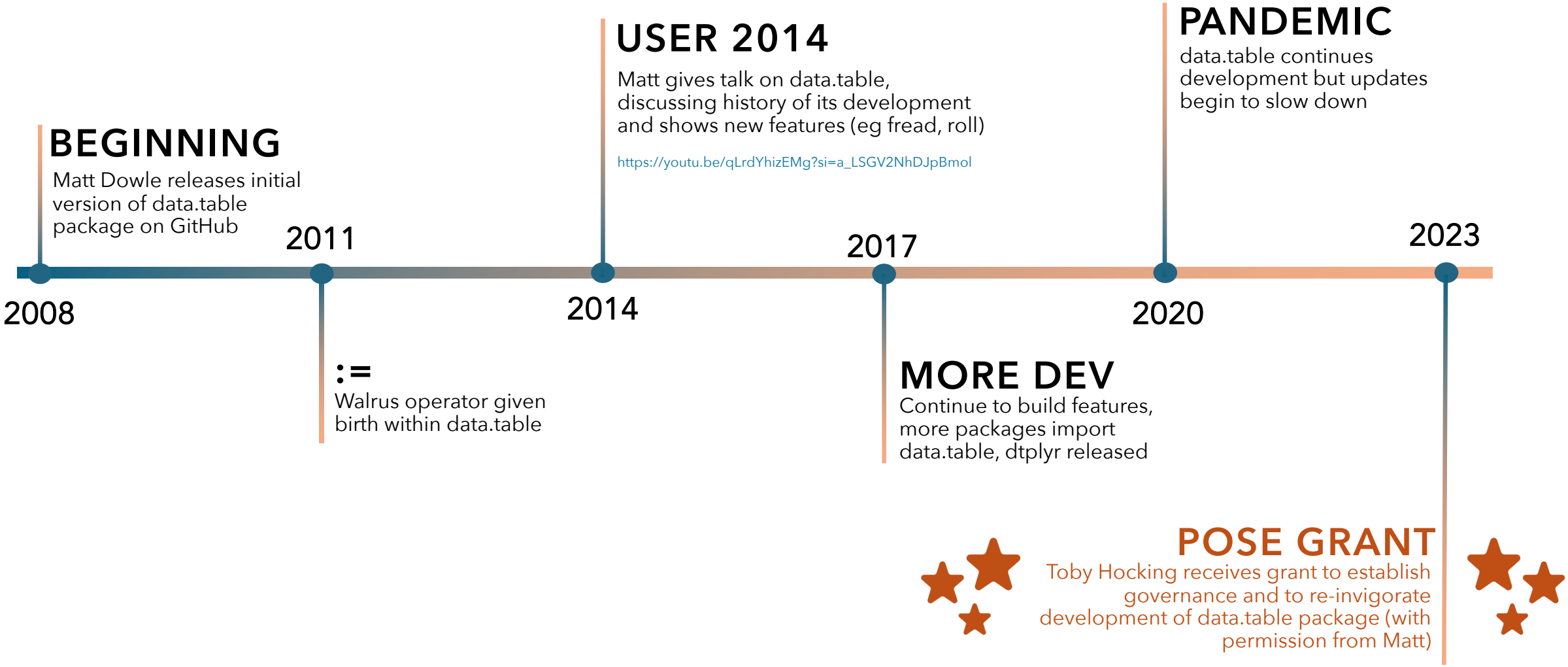
dt[i, j, by]

Concise syntax

```
dt[grp == “treatment”, new := mean(x), by = id]
dt[dt2, on = “id”]
dt[dt2, on = “id”, roll = TRUE]
dt[, .N, by = id]
```

PAST

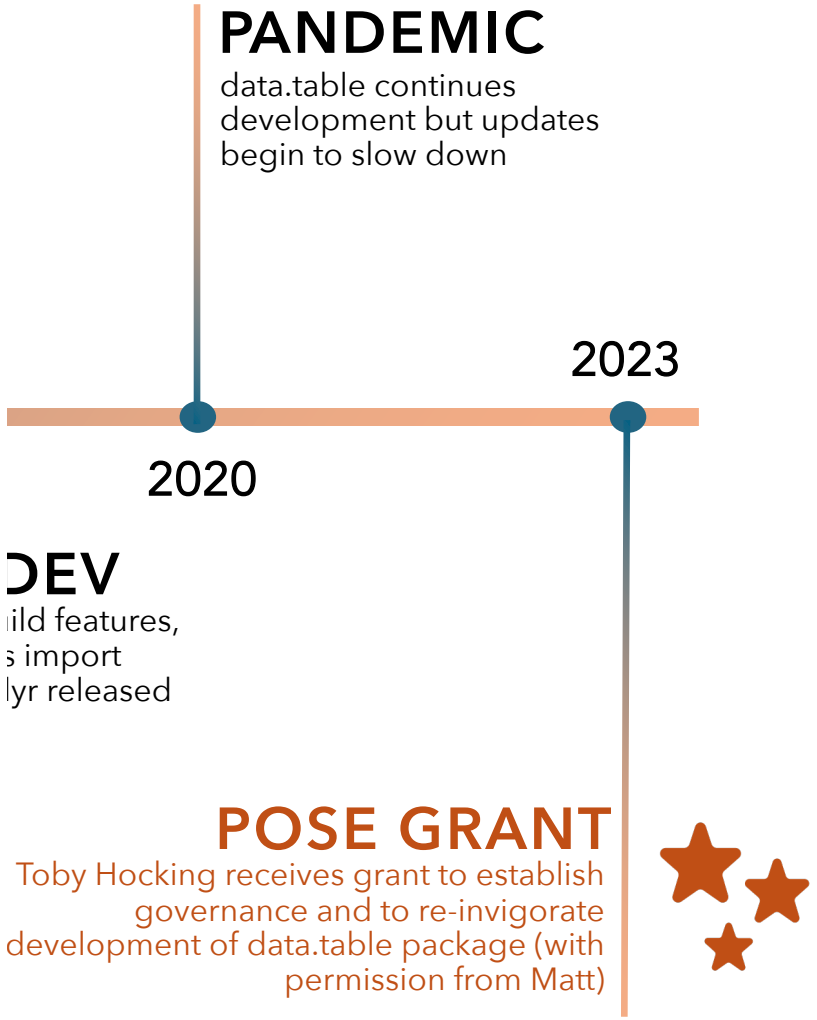

2008
BEGINNING
Matt Dowle releases initial version of data.table package on GitHub
2011
:=
Walrus operator given birth within data.table
2014
USER 2014
Matt gives talk on data.table, discussing history of its development and shows new features (eg fread, roll)
https://youtu.be/qLrdYhizEMg?si=a_LSGV2NhDJpBmol
2017
MORE DEV
Continue to build features, more packages import data.table, dtplyr released
2020
PANDEMIC
data.table continues development but updates begin to slow down
2023
POSE GRANT
Toby Hocking receives grant to establish governance and to re-invigorate development of data.table package (with permission from Matt)


BEGINNING
Matt Dowle releases initial version of data.table package on GitHub
2008
2011
:=
Walrus operator given birth within data.table
2014
USER 2014
Matt gives talk on data.table, discussing history of its development and shows new features (eg fread, roll)
https://youtu.be/qLrdYhizEMg?si=a_LSGV2NhDJpBmol
2017
MORE DEV
Continue to build features, more packages import data.table, dtplyr released
2020
PANDEMIC
data.table continues development but updates begin to slow down
2023
POSE GRANT
Toby Hocking receives grant to establish governance and to re-invigorate development of data.table package (with permission from Matt)

POSE: Phase II: Expanding the data.table ecosystem for efficient big data manipulation in R

Overview

More data is being systematically gathered and recorded than at any previous time in human history, thanks to the power of modern computer technology, which has revolutionized every scientific field of study (for example, ecology, physics, medicine, and finance). To extract patterns and knowledge from these increasingly large data sets, efficient software packages are required for data storage and analysis using limited computational resources. A leading, state-of-the-art example is `data.table`, which is free/open-source software for in-memory data manipulation/analysis, implemented as an R package with C code that is highly efficient in terms of both computation time and memory usage. `data.table` is a mature package that has been continuously developed since its initial release in 2006 and now has a substantial user base (for example, thousands of other R packages import and use functionality from `data.table`, including several R packages funded by NSF). However, the growth of `data.table` is limited by (1) its flat/informal leadership structure with only one author at the top who can approve new code contributions, (2) lack of documentation/translations and community standards for promoting diversity/inclusion, and (3) lack of infrastructure for systematic software testing. We propose to expand the open-source ecosystem of users, contributors, and developers of `data.table`, by fixing these problems. In particular, we propose to work on creating (1) a written governance document with a new hierarchical leadership structure, (2) new documentation materials for onboarding new users/contributors, including translations and community standards to encourage diversity/inclusion, and (3) new testing software and infrastructure. Furthermore, we plan to systematically evaluate the effects of our project on the `data.table` ecosystem by measuring changes to important metrics (number of unique contributors, diversity of contributors, number of dependent packages, etc). The result of this project will be a self-sustaining open-source ecosystem for `data.table`, which will allow it to grow into a more powerful data analysis tool that will be used by more people, and more diverse people, in the future.



PRESENT

TODAY

3 GOALS OF POSE GRANT

1. A written governance document with a new hierarchical leadership structure
2. New documentation materials for onboarding new users/contributors, including translations and community standards to encourage diversity/inclusion
3. New testing software and infrastructure

`data.table`

is a community-based open-source project that is sustainable and modern

PRESENT

TODAY

1 A written governance document with a new hierarchical leadership structure

https://github.com/Rdatatable/data.table/blob/master/GOVERNANCE.md

PRESENT

TODAY

1 A written governance document with a new hierarchical leadership structure

https://github.com/Rdatatable/data.table/blob/master/GOVERNANCE.md

A semi-democratic approach to dev

Can become any role in data.table by submitting PR and enough votes from the community (and no strong pushback)

Can help shape the development of the package

One aspect of the governance is the “what is possible for development” which can be updated

Decision-making processes

Definition of Consensus

Most decisions in the project happen by Consensus, which means that no active people (typically Reviewers and/or Committers) have expressed major blocking concerns, in a public discussion (typically in a GitHub issue or pull request). In Consensus, non-response by inactive members indicates tacit agreement.

Pull Requests

A pull request can be merged by any committer, if there is one approving review, and Consensus from active Reviewers and Committers.

- approving review must come from someone other than the author of the PR.
- approving review ideally comes from a reviewer of the affected files.
- approving review can and often will be by the committer who merges the PR.

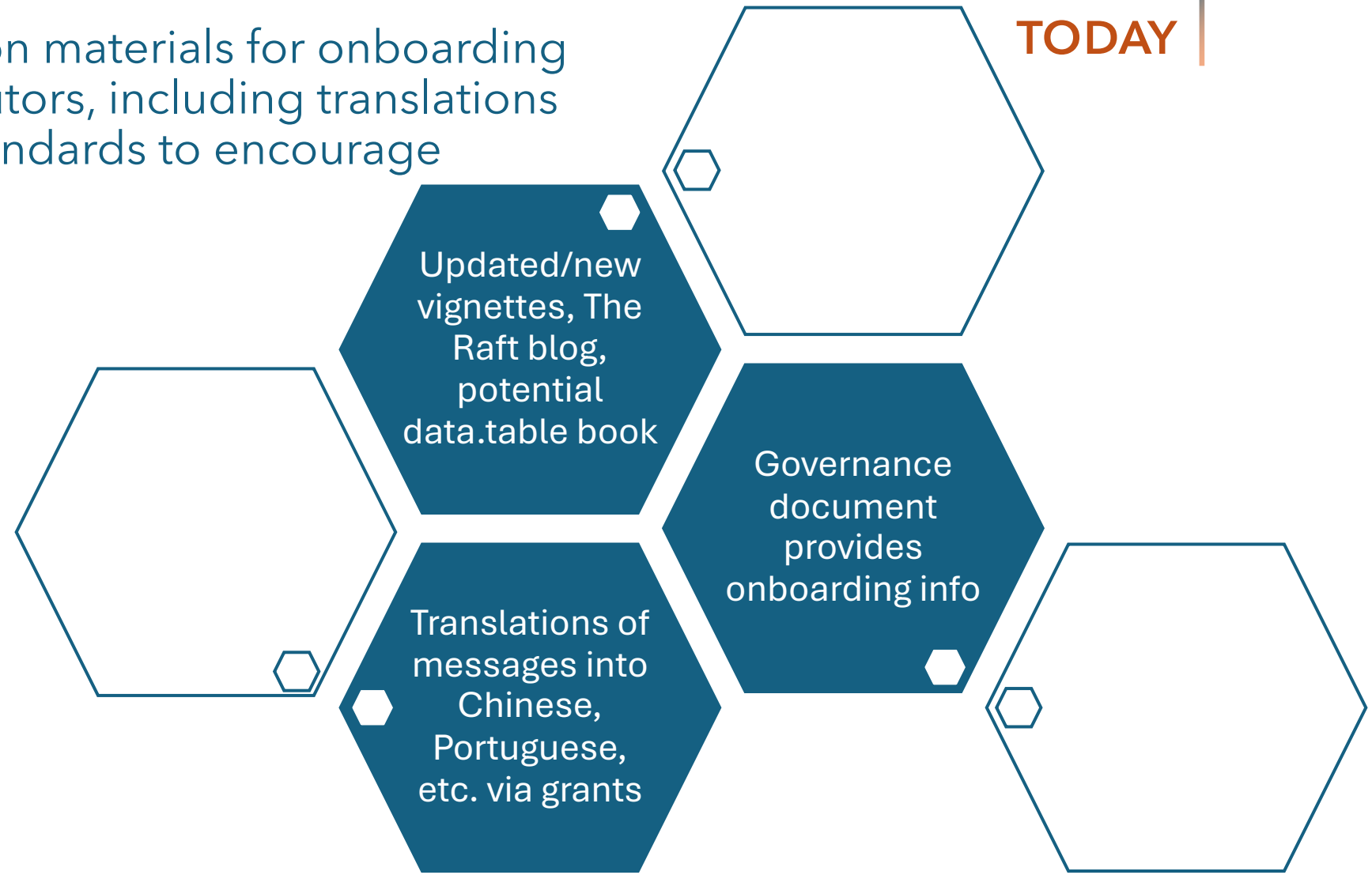
CRAN updates

- Regular CRAN releases should ideally occur twice per year, and can include new features.
- A hotfix/patch CRAN release should occur when CRAN asks for one, at which time the CRAN maintainer should post an issue on github, and ask others to help fix/prepare the release. It should not include new features.
- Both kinds of releases should be discussed in an issue, and the release should happen only if there is Consensus among active Reviewers and Committers.

PRESENT

TODAY

2 New documentation materials for onboarding new users/contributors, including translations and community standards to encourage diversity/inclusion



PRESENT

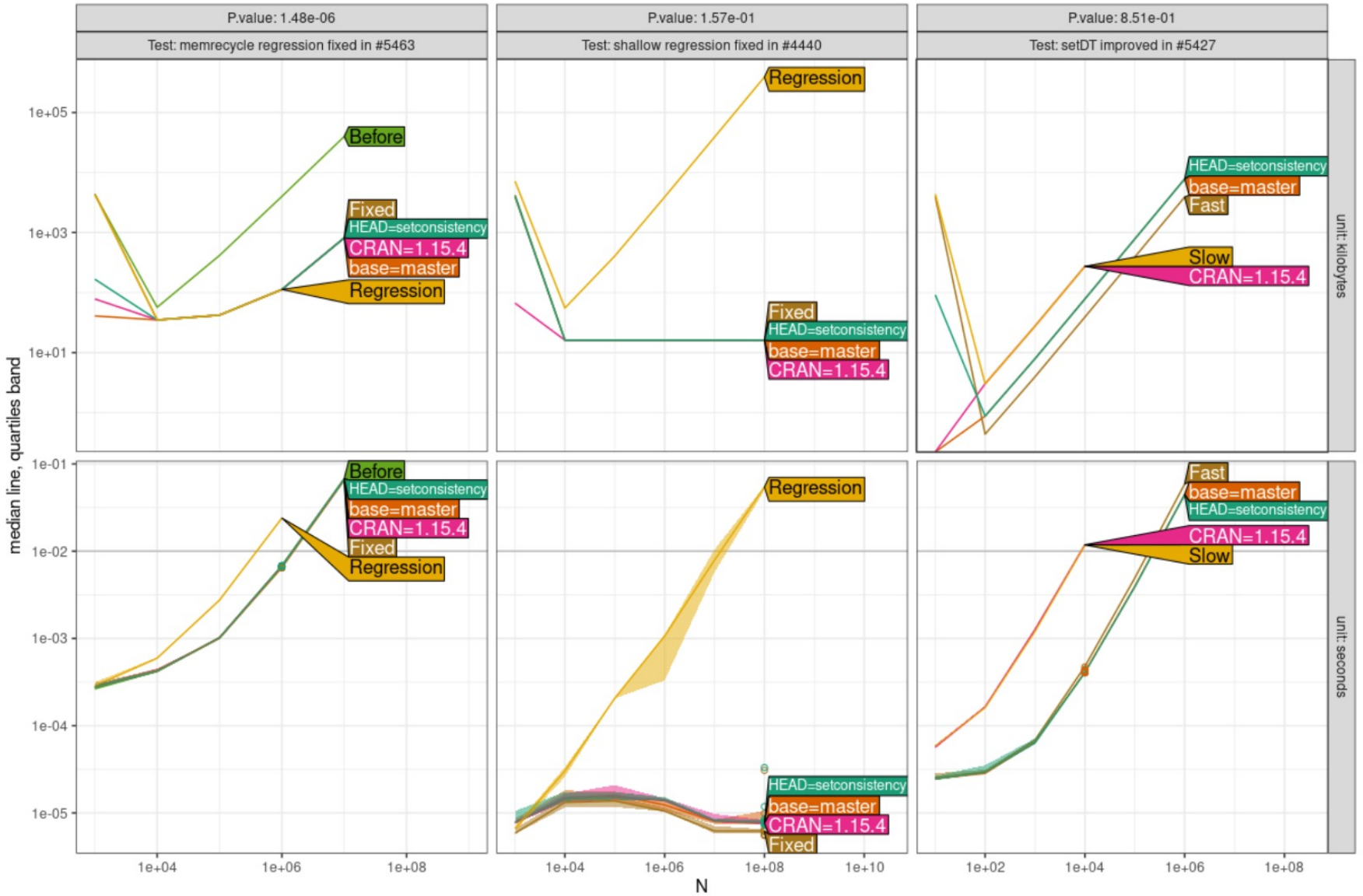
TODAY

3 New testing software and infrastructure

Beyond the normal unit-testing for the package (data.table has a lot!)

Benchmark how pull requests impact speed of the package

github-actions bot commented 3 hours ago



Generated via commit 931c81d

PRESENT

TODAY

NEW FEATURES

https://github.com/Rdatatable/data.table/blob/master/NEWS.md

Long awaited updates are happening now using the classic `data.table` principles

Using `dt[, names(.SD) := lapply(.SD, fx)]` now works, #795 -- one of our most-requested issues (see #3189). Thanks to @brodieG for the report, 20 or so others for chiming in, and @ColeMiller1 for PR.

Bug fixes and other notable improvements

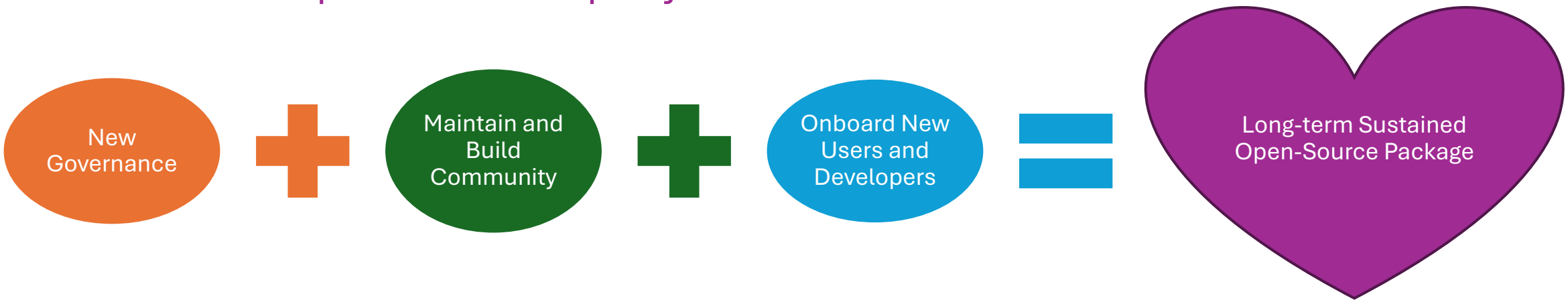
NOTES

1. `transform` method for data.table sped up substantially when creating new columns on large tables. Thanks to @OfekShilon for the report and PR. The implemented solution was proposed by @ColeMiller1.

2. The documentation for the `fill` argument in `rbind()` and `rbindlist()` now notes the expected behaviour for missing `list` columns when `fill=TRUE`, namely to use `NULL` (not `NA`), #4198. Thanks @sritchie73 for the proposal and fix.

FUTURE

The goal of this re-invigoration of data.table is to produce a sustainable open-source project



Beyond adding features, a few notable plans:

1. Establish pathways for funding (individual features or ongoing, eg NumFocus)
2. Build more infrastructure to help devs to parse the vast complexity of data.table package
3. More education on R's C APIs and how they can be used with compiled code

Tyson S. Barrett

Thanks to Matt Dowle, Arun Srinivasan, Michael Cherico, Jan Gorecki, Toby Hocking and the data.table team!

 t.barrett88@gmail.com

 tysonbarrett.com

 @healthandstats

 github.com/tysonstanley

 Slides will be on useR! 2024 website